

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 A PREDICTIVE CODING MODEL OF HIPPOCAMPO- NEOCORTICAL INTERACTIONS INVOLVED IN MEMORY REPLAY

006 **Anonymous authors**

007 Paper under double-blind review

010 ABSTRACT

013 The neocortex and the hippocampus are two complementary learning systems
014 which interact during memory construction and consolidation. The hippocampus
015 stores episodic memories coming from the neocortex passing through the entorhi-
016 nal cortex, and later replays them back to the neocortex to transform them into
017 semantic memory during memory consolidation. It is thought that memory replay
018 is a generative process, involved in imagining, because new episodes can also
019 be generated and instantiated in the neocortex. Here we present a computational
020 model of hippocampal-neocortical interactions based on a predictive coding net-
021 work with two hidden layers, which are mapped onto the visual cortex and the
022 entorhinal cortex. Improving on a previous implementation of this network, our
023 simulations provide a mechanistic account of memory replay in the neocortex.

024 1 INTRODUCTION

027 According to the complementary learning systems (CLS) theory, the neocortex is responsible for
028 semantic memory, that is the general knowledge that we have about the world, whereas hippocampus
029 stores episodic memories, which correspond to an individual’s emotional and sensory experiences
030 (Kumaran et al., 2016). For example, the experience of encountering a particularly odd-looking dog
031 (his look, bark, smell and the surprise you felt when seeing it) can be stored as an episodic memory
032 whereas the knowledge about what characterizes a typical dog is semantic memory.

033 After storage in the hippocampus, an episodic memory can be recalled from a corrupted version of it.
034 Furthermore, episodic memories are replayed during rest or sleep for memory consolidation. Mem-
035 ory replay corresponds to the spontaneous reactivation of the activity corresponding to an episodic
036 memory in the hippocampus (hippocampal replay) and its subsequent reinstatement in the neocor-
037 tex (cortical replay), so that it can gradually be integrated in the semantic memory of the neocortex.
038 This idea is supported by empirical evidence from rodent studies during spatial navigation, where
039 it was found that the rodent hippocampus generates sequences of activations during wakeful rest or
040 sleep that reflect past trajectories (Buzsáki, 2015). In machine learning, experience replay has been
041 shown to prevent catastrophic forgetting in a continual learning setting, where the learning of new
042 tasks interferes with the knowledge of previously learned tasks. It consists of continually storing
043 episodes in a memory buffer and replaying them when learning a new task.

044 It was later found that the internally generated hippocampal sequences are not merely replays of past
045 trajectories, but also include paths that were never experienced before (Kumaran et al., 2016). This
046 has prompted researchers from computational neuroscience and brain-inspired machine learning to
047 hypothesize that the hippocampus is a generative model and that memory replay is a generative
048 process, referred to as generative replay (Stoianov et al., 2022; van de Ven et al., 2020). Moreover,
049 generative replay has been shown to improve the performance of reinforcement learning agents over
experience replay Wang et al. (2025).

050 Despite the importance of memory replay for continual learning, its implementation in the neocortex
051 and the hippocampus is not well understood. Because of the similarity of architecture in all its
052 areas, researchers hypothesized that a common algorithm underlies computations in the neocortex
053 (Friston, 2003; Hawkins et al., 2019). Principles of organization in the neocortex have emerged from
empirical studies in the visual cortex, which have shown that this region is arranged hierarchically,

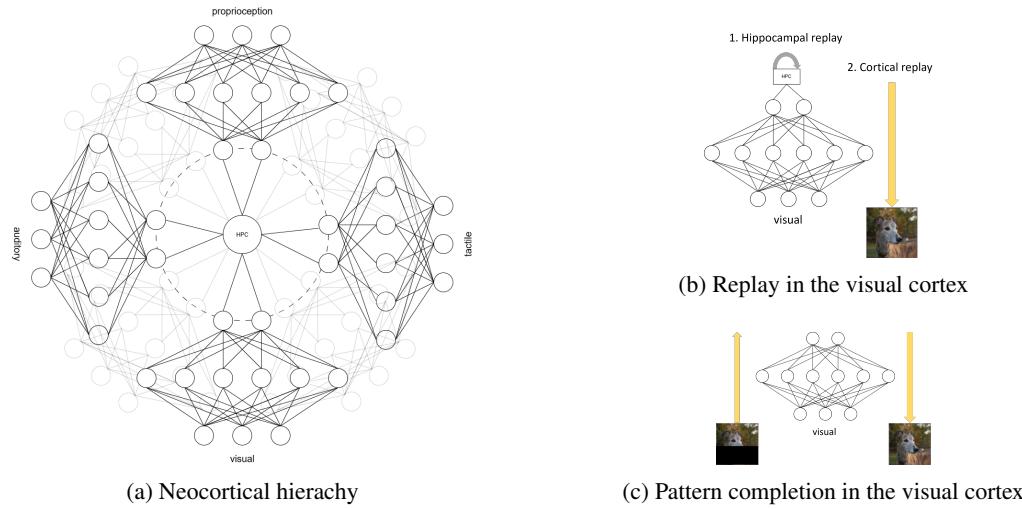


Figure 1: (a) The high-level representations in neocortical areas corresponding to different sensory modalities are combined in the entorhinal cortex (represented as the dotted circle) which is the input of the hippocampus (HPC), located at the apex of the neocortical hierarchy. (b) After storage of an episodic memory in the hippocampus, it can be replayed in the hippocampus (hippocampal replay) before being reinstated in the neocortex, including the visual cortex (cortical replay). (c) The visual cortex can complete an image based on semantic memory. In this illustration, the completion performance is perfect, as the recalled image is exactly the original image (as in Salvatori et al. (2021)), but it is not the case in reality (as explained in the main text).

with forward connections from lower to higher areas, and backward connections from higher to lower areas (Friston, 2003). At the apex of the neocortical hierarchy, the hippocampus receives input from the entorhinal cortex, which combines representations from different high-level neocortical areas of different sensory modalities (Barron et al., 2020), as illustrated in Figure 1a. Therefore, hippocampal replay could drive neocortical activity using backward connections from the entorhinal cortex to the neocortical hierarchy, as illustrated for the visual cortex in Figure 1b.

Rooted in studies of the visual cortex, predictive coding has been proposed in computational neuroscience by Rao & Ballard (1999), and later extended by Friston (2003) as a general theory of cortical computation, which maps well to the neocortex in terms of architecture and information processing. Recently, Fontaine & Alexandre (2025b) investigated the role of the neocortex in semantic and episodic memory using a predictive coding network (PCN). They reproduced the result of Salvatori et al. (2021) that PCNs can store training images as memories (as illustrated in 1c), but showed that this is done by overfitting the network to a few training images. When the network is trained on more images, it generalizes better and is able to complete corrupted versions of training images based on semantic memory, but without recalling the details of the specific training images, supporting the CLS view that the neocortex is responsible for semantic memory. Even though the neocortex might not be responsible for episodic memory like the hippocampus, it supports episodic memory by allowing the episodic memories replayed by the hippocampus to be reinstated during cortical replay. Fontaine & Alexandre (2025b) modelled experience replay in a PCN, using two classes of MNIST digits, but the replayed images were found to be blurry, and the representations at the top of the hierarchy were found to be overlapping at the boundary of the two classes. In this paper, we tackle these limitations by accurately tuning the number of hidden units and stabilizing the convergence of the model on the full MNIST dataset using a learning rate scheduler. We found that adding more neurons in the second hidden layer allows the experience replays to be more accurate, by increasing the linear separability of representations corresponding to different classes at the top of the hierarchy, but at the cost of reconstruction and pattern completion performance. Furthermore, we extend the model to generative replay, by proposing a generative form of hippocampal replay.

108

2 RELATED WORKS

110 **Models of memory replay.** Other works captured the computation of experience replay and generative replay using deep learning models, but without being faithful to the architecture and information processing of the corresponding structures in the brain. (Spens & Burgess, 2024) proposed a model 111 of memory consolidation consisting of experience replay, in which the neocortex is modelled as a VAE, whereas the hippocampus is modelled as a Modern Hopfield Network (Krotov & Hopfield, 112 2016) which replays episodic memories to the input of the VAE. In Stoianov et al. (2022), the hippocampus 113 itself is modelled as a hierarchical generative model supporting generative replay.

114 **PCNs as models of the neocortex.** Since the work of Friston (2003), various papers investigated the 115 idea that predictive coding could underly information processing in the neocortex. Brucklacher et al. 116 (2023) proposed to study the representations learned by a PCN with two hidden layers and showed 117 that the representations in the highest area are object-invariant when trained on sequences of continuously 118 transformed images. In addition, they show that top-down reconstruction of inputs from latent 119 variables when blanking out the input becomes less accurate in higher areas, suggesting that higher 120 areas encode reduced information such as object identity. In our model, memory replay is modelled 121 without blanking out the input, in a biologically plausible manner, as memory replay can occur while 122 the brain is exposed to sensory input. Salvatori et al. (2021) showed that PCNs outperform other 123 models in auto-associative memory (AM) and suggested based on Barron et al. (2020) that the top 124 layer of their PCN could correspond to the hippocampus. Tang et al. (2023) extended their model by 125 adding a recurrent one-layer PCN to the top of a hierarchical PCN modelling the neocortex and Li 126 et al. (2025) showed that PCNs also detect novelty at different levels of abstraction in the hierarchy. 127 The architecture and the inference and learning rules of our model are similar to that of Salvatori 128 et al. (2021) as both models are based on the model by Friston (2003). The main difference is that 129 our work models memory replay and not auto-associative memory, which is a function associated 130 to the hippocampus rather than the neocortex according to Fontaine & Alexandre (2025b). However, 131 the pattern completion performance of our model will be evaluated on unseen images, to tune 132 the size of the top level. Indeed, even though the neocortex likely cannot recall details of episodic 133 memories, it can still complete corrupted patterns based on semantic memory. Another difference 134 with this line of work is that the top level of our PCN corresponds to the entorhinal cortex, and 135 not the hippocampus. Indeed, as the hippocampus is able to perform one-shot storage of episodic 136 memories, we believe that it cannot be fully modelled using a PCN.

137 **Image generation in PCN.** Other works studied the generation of inputs in predictive coding 138 networks (PCN), but without specifying how it can be implemented in the neocortex in the presence of 139 sensory inputs, or its relation to the hippocampus. On the one hand, Oliviers et al. (2024) proposed 140 Monte Carlo predictive coding for learning probability distributions of sensory inputs, arguing that 141 classical predictive coding demonstrated limited performance in generative tasks. On the other hand, 142 preliminary work by Millidge (2019) showed that a PCN with one hidden layer can be used to generate 143 inputs by sampling points close to the training data in the latent space, even though the generated 144 samples were blurry and the authors did not describe their method for sampling. Ororbia & Kifer 145 (2022) proposed to generate images from an extended version of PCN with ancestral sampling, but 146 only three images per class are shown. In addition, using the model proposed by Friston (2003), the 147 authors only show nearest neighbor samples that match an original data point for each class, leaving 148 aside a large part of the image space which is covered by the generative model.

149

3 PREDICTIVE CODING

150 Predictive coding networks (PCN) are based on hierarchical generative models with L layers, of the 151 type

$$152 \quad \forall l \in \{0, 1, \dots, L-1\}, p_{\theta_l}(\mathbf{h}_l | \mathbf{h}_{l+1}) = \mathcal{N}(\mathbf{h}_l; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \boldsymbol{\mu}_l = f_l(\boldsymbol{\mu}_{l+1}; \boldsymbol{\theta}_l) \\ 153 \quad p_{\theta_L}(\mathbf{h}_L) = \mathcal{N}(\mathbf{h}_L; \boldsymbol{\mu}_L, \boldsymbol{\Sigma}_L), \boldsymbol{\mu}_L = \boldsymbol{\theta}_L$$

154 where $\mathbf{h} = (\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_L)$ are the states and $\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_L)$ are the parameters. Level 0 is 155 the input level, so \mathbf{h}_0 is the input state and $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L$ are the latent states. In practice, we use 156 $\boldsymbol{\mu}_L = \mathbf{0}$ and $\boldsymbol{\Sigma}_l = \mathbf{I}$ for all $l \in \{0, 1, \dots, L\}$. Therefore, the prior on the latent state of level L is a 157 centered isotropic multivariate Gaussian $p_{\theta_L}(\mathbf{h}_L) = \mathcal{N}(\mathbf{h}_L; \mathbf{0}, \mathbf{I})$ with no parameter.

162 Recognition is assumed to be deterministic, such that for an input \mathbf{x} ,
 163

$$164 \quad q_{\phi}(\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L \mid \mathbf{x}) = \prod_{l=1}^L \delta(\mathbf{h}_l - \phi_l)$$

166 where $\phi = (\phi_1, \phi_2, \dots, \phi_L)$ is an estimate of the latent states $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L$ corresponding to the
 167 input \mathbf{x} .
 168

169 Inference of the latent states for an input \mathbf{x} results from the minimization of a lower bound to the
 170 negative likelihood, called variational free energy
 171

$$\begin{aligned} \mathcal{L}(\theta, \phi; \mathbf{x}) &= -\mathbb{E}_{q_{\phi}(\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L \mid \mathbf{x})} [\log p_{\theta}(\mathbf{x}, \mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_L)] \\ &= -\log p_{\theta}(\mathbf{x}, \phi_1, \phi_2, \dots, \phi_L) \\ &= -\log p_{\theta_0}(\mathbf{x} \mid \phi_1) - \log p_{\theta_1}(\phi_1 \mid \phi_2) - \dots - \log p_{\theta_{L-1}}(\phi_{L-1} \mid \phi_L) - \log p(\phi_L) \\ &= \frac{1}{2} \sum_{l=0}^{L-1} [\xi_l^T \xi_l + \log |\Sigma_l|] - \log p(\phi_0) + \text{constant} \end{aligned}$$

172 where taking the logarithm of a Gaussian distribution results in a quantity
 173

$$\xi_l = \Sigma_l^{-\frac{1}{2}} (\phi_l - f_l(\phi_{l+1}; \theta_l)). \quad (1)$$

174 which can be seen as a prediction error for layer l . Thus, the variational free energy corresponds to
 175 the sum of prediction errors in all layers, and PCNs learn hierarchical predictive representations of
 176 the input.
 177

178 When presented with an input \mathbf{x} , the latent states are updated according to
 179

$$\dot{\phi}_l = -\nabla_{\phi_l} \mathcal{L}(\theta, \phi; \mathbf{x}) = -\frac{\partial \xi_{l-1}^T}{\partial \phi_l} \xi_{l-1} - \frac{\partial \xi_l^T}{\partial \phi_l} \xi_l \quad (2)$$

180 until the variational free energy is minimized. In practice, we only update the latent states T times
 181 during training.
 182

183 Similarly, learning of the parameters $\theta = (\theta_0, \theta_2, \dots, \theta_{L-1})$ corresponds to the minimization of
 184 the variational free energy $F(\theta) = \mathbb{E}_{p(\mathbf{x})} [\mathcal{L}(\theta, \phi; \mathbf{x})]$. After several image presentations, the
 185 parameters are updated once following
 186

$$\forall l \in \{1, 2, \dots, L\}, \dot{\theta}_l = -\nabla_{\theta_l} F = -\mathbb{E}_{p(\mathbf{x})} \left[\frac{\partial \xi_l^T}{\partial \theta_l} \xi_l \right]. \quad (3)$$

187 This algorithm can be implemented in a neural network, hence the name PCN, with only local
 188 computations for inference and learning. In a PCN, each level l consists of two types of neurons,
 189 with activity ϕ_l and ξ_l respectively. When mapped to the neocortical hierarchy, level $l+1$ is the
 190 level above l , with level 0 at the bottom and level L at the top. From equation 1, it can be seen that
 191 neurons ξ_l compute the prediction errors, based on lateral connections with neurons ϕ_l at the same
 192 level and inhibitory feedback connections with neurons ϕ_{l+1} in the level above, which provide the
 193 predictions. Equation 2 shows that neurons ϕ_l receive connections from error neurons in the same
 194 level ξ_l and the level below ξ_{l-1} . In addition, it can be seen that equation 3 corresponds to Hebbian
 195 learning, as shown in the next section.
 196

197 While it is standard in the predictive coding literature to use fixed covariances $\Sigma_l = \mathbf{I}$, it was
 198 proposed that the inverse covariance, called precision, Σ_l^{-1} is predicted by higher layers and
 199 mediates attention (Feldman & Friston, 2010). Indeed, if precision Σ_l^{-1} is low, prediction error
 200 $\xi_l = \Sigma_l^{-\frac{1}{2}} (\phi_l - f_l(\phi_{l+1}; \theta_l))$ is low, and will not influence the update of neuron activities and
 201 weights. As explained by Li (2023) based on Clark (2016), the precision parameter controls the
 202 degree to which the brain attends to the external input or to the internal prediction, determining
 203 whether it is performing perception or imagination.
 204

205 4 METHODS

206 Building upon Fontaine & Alexandre (2025a), we propose a predictive coding model of the visual
 207 cortex and show that it learns hierarchical predictive representations of MNIST images, that support
 208 memory replay.
 209

216 4.1 MODEL
217218 Our model is a PCN with $L = 2$ layers
219

220
$$\begin{aligned} p_{\theta_0}(\mathbf{h}_0 \mid \mathbf{h}_1) &= \mathcal{N}(\mathbf{h}_0; \mathbf{W}_0 \mathbf{h}_1, \mathbf{I}) \\ p_{\theta_1}(\mathbf{h}_1 \mid \mathbf{h}_2) &= \mathcal{N}(\mathbf{h}_1; f(\mathbf{W}_1 \mathbf{h}_2 + \mathbf{b}_1), \mathbf{I}) \\ p(\mathbf{h}_2) &= \mathcal{N}(\mathbf{h}_2; \mathbf{0}, \mathbf{I}) \end{aligned}$$

221

222 where $\theta_0 = \mathbf{W}_0$ and $\theta_1 = (\mathbf{W}_1, \mathbf{b}_1)$. Indeed, the bias and non-linearity are not required in the input
223 layer, because an image can be represented as a linear combination of basis functions (Olshausen &
224 Field, 1996).
225226 The variational free energy for an input \mathbf{x} is
227

228
$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \frac{1}{2} \xi_0^\top \xi_0 + \frac{1}{2} \xi_1^\top \xi_1 + \frac{1}{2} \xi_2^\top \xi_2 + \text{constant}$$

229

230 where
231

232
$$\xi_0 = \mathbf{x} - \mathbf{W}_0 \phi_1 \quad (4)$$

233

234
$$\xi_1 = \phi_1 - f(\mathbf{W}_1 \phi_2 + \mathbf{b}_1) \quad (5)$$

235

236
$$\xi_2 = \phi_2. \quad (6)$$

237

238 4.2 TRAINING ALGORITHM
239240 Let us consider a dataset $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$ of N i.i.d. samples of a continuous variable \mathbf{x} . The log
241 likelihood can be written $\log p_{\theta}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) = \sum_{i=1}^N \log p_{\theta}(\mathbf{x}^{(i)})$. Therefore, the variational
242 free energy of the dataset \mathbf{X} is
243

244
$$\mathcal{L}(\theta; \mathbf{X}) = \sum_{i=1}^N \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}),$$

245

246 which can be estimated based on minibatches
247

248
$$\mathcal{L}(\theta; \mathbf{X}) \approx \mathcal{L}^M(\theta; \mathbf{X}^M) = \frac{N}{M} \sum_{i=1}^M \mathcal{L}(\theta, \phi^{(i)}; \mathbf{x}^{(i)}) \quad (7)$$

249

250 where the minibatch $\mathbf{X}^M = \{\mathbf{x}^{(i)}\}_{i=1}^M$ is randomly drawn from the dataset \mathbf{X} .
251252 Given a minibatch \mathbf{X}^M , the latent states $\phi_0^{(i)}$ and $\phi_1^{(i)}$ are updated during T iterations for each
253 datapoint $\mathbf{x}^{(i)}$ to minimize the variational free energy $\mathcal{L}(\theta, \phi^{(i)}; \mathbf{x}^{(i)})$. The update rules for the two
254 layers in our model are
255

256
$$\Delta \phi_1 = \alpha(\mathbf{W}_0^\top \xi_0 - \xi_1) \quad (8)$$

257

258
$$\Delta \phi_2 = \alpha(\mathbf{W}_1^\top \text{diag}[f'(\mathbf{W}_1 \phi_2 + \mathbf{b}_1)] \xi_1 - \xi_2). \quad (9)$$

259

260 where α is the inference rate.
261262 Then, the parameters $\theta_0 = \mathbf{W}_0$ and $\theta_1 = (\mathbf{W}_1, \mathbf{b}_1)$ are updated once to minimize the estimate
263 $\mathcal{L}^M(\theta; \mathbf{X}^M)$ given in equation 7. Thus, the learning rules can be calculated from the sum of gradi-
264 ents $\sum_{i=1}^M \nabla_{\theta_l} \mathcal{L}(\theta; \mathbf{x}^{(i)})$
265

266
$$\begin{aligned} \Delta \mathbf{W}_0 &= \beta \sum_{i=1}^M \xi_0^{(i)} \phi_1^{(i)\top} \\ \Delta \mathbf{W}_1 &= \beta \sum_{i=1}^M \left[\xi_1^{(i)} \odot f'(\mathbf{W}_1 \phi_2^{(i)} + \mathbf{b}_1) \right] \phi_2^{(i)\top} \\ \Delta \mathbf{b}_1 &= \beta \sum_{i=1}^M \xi_1^{(i)} \odot f'(\mathbf{W}_1 \phi_2^{(i)} + \mathbf{b}_1). \end{aligned}$$

267

268 where \odot is the element-wise product and β is the learning rate. We train the model on a dataset \mathbf{X}
269 for multiple epochs, and reduce the learning rate by a multiplicative factor $\gamma < 1$ at each epoch of
270 training to prevent instability issues. Details can be found in section A.1 of the appendix.
271

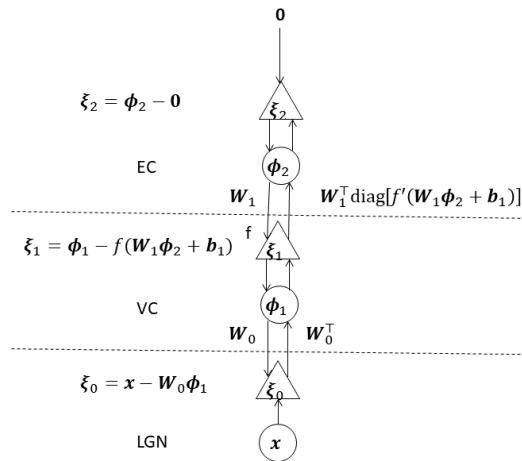


Figure 2: PCN with 3 levels mapped to lateral geniculate nucleus (LGN), visual cortex (VC) and entorhinal cortex (EC). Each node corresponds to multiple neurons. Circle nodes correspond to units ϕ_l and triangle nodes correspond to error units ξ_l . Each connection between nodes corresponds to a fully connected network, with excitatory connections in the upward direction and inhibitory connections in the downward direction. We consider the upmost prediction to be 0, to have a standard normal distribution as prior.

4.3 MAPPING TO THE BRAIN

Figure 2 shows how this algorithm can be implemented in a neural network mapped onto the visual pathway of the brain, with only local computations. The three levels in the network correspond, from bottom to top, to the lateral geniculate nucleus (LGN) in thalamus (and not the retina because it doesn't receive feedback connections from LGN), the visual cortex (VC) and the entorhinal cortex (EC). This mapping allows us to study memory replay after training the model with algorithm 1.

At inference time, the latent states are updated until convergence or until a maximum number of iterations T_{\max} is reached in order to study the converged representations. Convergence occurs when the relative change in the norm of the latent state ϕ_l is smaller than a threshold ϵ , i.e.

$$\frac{\|\nabla_{\phi_l} \mathcal{L}(\theta, \phi; x)\|}{\|\phi_l\|} < \epsilon.$$

During perception, the network is driven by the input image in LGN. While the LGN is set to the image, the VC and EC converge to hierarchical predictive representations of the image following the inference rules 8 and 9 respectively. The representation in VC is predictive of the image, as the prediction $W_0 \phi_1$ is a reconstruction of the image. The representations in EC can be stored by the hippocampus (not explicitly modelled) and later replayed during experience replay.

During memory replay, hippocampal replay first outputs the EC representation of an image stored by the hippocampus in the case of experience replay or a sample generated in the latent space of EC in the case of generative replay (as described in the next paragraph). Then, during cortical replay, the network is driven by the representation in EC obtained from hippocampal replay. While setting the EC layer to the corresponding representation, the VC converges to the replayed representation following the inference rule 8. As the network can also be presented with an input in LGN during memory replay, the representation in VC is protected from ascending input in LGN by setting the precision Σ_0^{-1} in the LGN to 0, preventing the prediction errors in LGN to influence the activity in VC. In this way, attention is focused on the representation in EC, and not on the current input. Then, the prediction $W_0 \phi_1$ based on the replayed representation ϕ_1 in VC corresponds to the replayed image.

In generative replay, sampling of the latent space of EC is class-conditioned. Indeed, we fit a multivariate Gaussian distribution to each class in the latent space of EC, by estimating the mean and

324 covariance of training samples in each class. Then, samples from a given class can be generated by
 325 sampling the corresponding Gaussian distribution.
 326

327 During perception, the network can also perform pattern completion when presented with a cor-
 328 rupted (noisy or incomplete) input. Following Salvatori et al. (2021), while the LGN layer is ini-
 329 tialized to the corrupted input, the VC and EC converge following the same dynamics as in regular
 330 perception (equations 8 and 9), whereas the corrupted part of LGN converges following to the infer-
 331 ence rule

$$\Delta\phi_0 = -\alpha\phi_0.$$

332 obtained by calculating the gradient of the variational free energy of the input with respect to the
 333 corrupted part of LGN. Then, the prediction $W_0\phi_1$ based on the recalled representation ϕ_1 in VC
 334 corresponds to the recalled image.
 335

336 4.4 EXPERIMENTS

337 The model is trained on the MNIST dataset, which contains images of handwritten digits from 0
 338 to 9. The original training set of 60,000 images is split into a training set of size 50,000 and a
 339 validation set of size 10,000. The validation set is used to evaluate the model during training and
 340 hyperparameter tuning. After training, we evaluate the model on the original test set of 10,000
 341 images.
 342

343 Hyperparameter values are chosen based on empirical trials informed by the predictive coding liter-
 344 ature and summarized in Table 1 in the appendix. In addition, the number of hidden units in level
 345 1 is obtained by minimizing the variational free energy on the validation set for a PCN with $L = 1$
 346 using grid search, as shown in Figure 7 in the appendix. The choice of the number of hidden units
 347 in level 2 is more complex and results from a trade-off between different metrics.

348 In our simulations, we study the influence of the number of hidden units in level 2 on the learned
 349 representations and on memory replay, both quantitatively and qualitatively. Quantitatively, we
 350 evaluate the predictive performance of the model, the quality of the experience replays and the
 351 pattern completion performance using the reconstruction, replay and completion errors, based on
 352 the mean squared error (MSE). The MSE between two flattened images \mathbf{x} and $\hat{\mathbf{x}}$ is

$$353 \text{MSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N_{\text{pixels}}} \sum_{i=1}^{N_{\text{pixels}}} (x_i - \hat{x}_i).$$

356 The reconstruction and completion errors are computed between the original and reconstructed im-
 357 ages and between the original and recalled images respectively, averaged over the validation set,
 358 whereas the replay error is computed between the original and replayed images, averaged over the
 359 training set. In addition, we evaluate the linear separability of the latent manifolds corresponding to
 360 the different classes in level 2 of the model using the classification accuracy of a simple multinomial
 361 logistic regression. On the qualitative side, we examine examples of reconstructions of images from
 362 the validation set, as well as examples of experience replay corresponding to images from the train-
 363 ing set. Additionnally, we show examples of images generated by generative replay and visualize
 364 their hierarchical representations.
 365

366 5 RESULTS

367 The reconstruction error, completion error, replay error and classification accuracy are plotted
 368 against the number of hidden units in level 2 in Figure 3. These measures are evaluated after train-
 369 ing the model until convergence of all layers, as shown in Figure 8 in the appendix. One the one
 370 hand, the left plot of Figure 3 shows that the second hidden layer in our model does not improve the
 371 predictive power of its representations. On the contrary, the reconstruction error increases with the
 372 number of units in the second hidden level, and the same model with the top layer removed has a
 373 lower reconstruction error (shown as the blue dotted line) than any of the models with two hidden
 374 layers. Similarly, the completion error increases with the number of units in the second hidden level,
 375 but it is always lower than for a model with only one hidden layer, suggesting that adding a second
 376 layer is still beneficial to the completion performance, contrary to the reconstruction performance.
 377 On the other hand, the right plot of Figure 3 shows that the quality of replay and the linear separa-
 378 bility of the classes in the second hidden level increases with the number of units in that level.

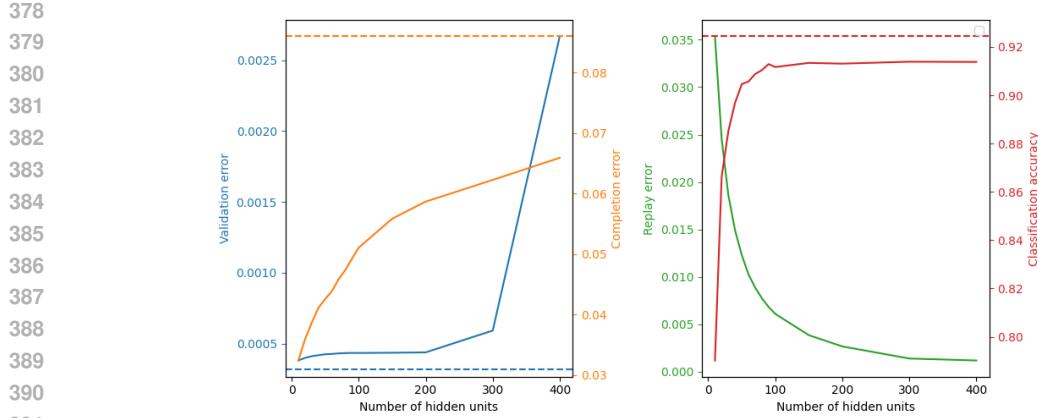


Figure 3: Quantitative evaluation of our model according to different metrics, depending on the number of hidden units in level 2. The dotted lines correspond to the same model with $L = 1$.

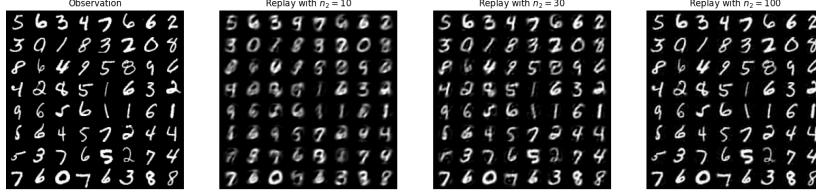


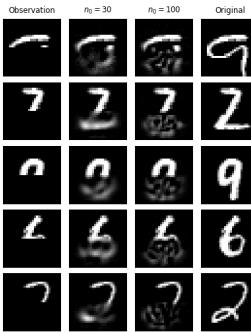
Figure 4: The same minibatch of images from the training set (shown in Observation) is replayed in PCNs with different widths n_2 at the top level.

Therefore, choosing the number of units in level 2 based on these metrics is not straightforward, and we will turn to the qualitative evaluation.

In Figure 4, we visualize examples of experience replay obtained with models of different widths. As shown in Fontaine & Alexandre (2025a), setting the width of the top level to the number of classes (i.e. $n_2 = 10$) results in replayed images that are blurry and that do not retain the details of the original images. This issue is solved by increasing the width of the top level to 30. Increasing it further to 100 improves the sharpness of the replayed images, but at the cost of the completion performance. Indeed, Figure 5 shows that while a width of 30 enables the network to semantically complete the bottom half of images taken from the validation set that were masked, a higher width of 100 makes the completion uninformative of the classes of the masked digits.

However, visual inspection of the reconstructions of images from the validation set and generative replays does not differentiate the models with different widths. Indeed, the difference in reconstruction errors between the different models is imperceptible in the reconstructed images. Similarly, the latent spaces and the quality of the images generated by replay are similar in the different models, despite the difference in replay error and classification accuracy. Thus, the trade-off between the replay fidelity and completion performance leads us to choose a width of 30.

Examples of reconstructions of images from the validation set and replayed images obtained by generative replay for a model with $n_2 = 30$ units in level 2 are shown in Figure 6. It can be seen from Figure 6a that the representations learned by the model are perfectly predictive of images it has never seen during training. In Figure 6c, we plotted hierarchical representations obtained with generative replay, over hierarchical representations inferred from images of the validation set. The representations of the different classes are well separated in all three levels, including the top level which was found to have overlapping clusters in Fontaine & Alexandre (2025a), and the representations generated by generative replay mostly fall within the right clusters in all levels. This is confirmed



432
 433
 434
 435
 436
 437
 438
 439
 440
 441
 442
 443
 444
 445
 446
 447
 448 Figure 5: Five images from the validation set (shown in the Original column) were masked (as
 449 shown in the Observation column) and completed by a PCN with $n_2 \in \{30, 100\}$ neurons in the top
 450 level.

448 by looking at the generated images in Figure 6b. Most of the images generated for each class are
 449 realistic examples of their classes, even though some of them are blurry.

451 6 DISCUSSION

452
 453 We proposed a model of hippocampo-neocortical interactions involved in memory replay using a
 454 PCN. Our work shows that predictive coding accounts for cortical replay, i.e. the reinstatement in
 455 the neocortex of an episodic memory replayed or generated in the hippocampus. To this purpose,
 456 we modelled hippocampal replay in a minimal way, both in experience replay and generative re-
 457 play, by mapping the top level of the PCN to the output of the hippocampus, the entorhinal cortex.
 458 Some of the images generated by our model were found to be out-of-distribution, probably because
 459 the simple Gaussian distribution we used does not capture the complex, non-linear geometry of the
 460 latent manifolds. This issue could be solved using a Riemannian metric (Arvanitidis et al., 2021).
 461 However, to provide a more complete account of memory replay, future work should aim at mod-
 462 ellelling the hippocampal formation with its different components to understand hippocampal replay
 463 mechanistically. In this way, realistic in-distribution samples will naturally be generated thanks to
 464 the learned connection between the entorhinal cortex and the hippocampus. In the hippocampus,
 465 generative replay should encompass experience replay as a generative process which samples both
 466 existing episodic memories and imagined ones.

467 Our work also contributes to understanding hierarchical representations in PCNs. It reveals on the
 468 one hand that adding more neurons in the top level of the network improves the fidelity of experience
 469 replay and the linear separability of the representations corresponding to different classes in the
 470 top level. Indeed, expanding the dimensionality of the activity space of patterns increases their
 471 separability (Cayco-Gajic & Silver, 2019), and the increased separability could lead to better replay,
 472 as there is less interference between patterns. However, our study also shows that the width of the
 473 top level is detrimental to the model’s predictive and completion performance on unseen images. We
 474 also found that adding a second hidden layer is beneficial to pattern completion, but not to image
 475 reconstruction. The negative result about reconstruction can be interpreted in the light of efficient
 476 coding: any image can be described by the linear combination of a set of basis functions (Olshausen
 477 & Field, 1996). In our model, these basis functions correspond to the weights of the bottom layer,
 478 between VC and LGN, which enable for the prediction of images in LGN. As we have tuned the
 479 number of neurons in VC to minimize the reconstruction error on the validation set (see Figure 7 in
 480 the appendix), we have found such a set of basis functions for the MNIST dataset. Therefore, adding
 481 a second layer can only decrease the validation error. These results suggest that layers higher in the
 482 neocortical hierarchy have a role in pattern completion, but not in reconstruction. Furthermore, the
 483 work of (Brucklacher et al., 2023) and (Li et al., 2025) indicate that higher areas in PCNs encode
 484 object identity. However, we find that the classification accuracy of our model is higher if we remove
 485 the top hidden layer (shown as the red dotted line in Figure 3). The role of depth will be investigated
 486 in future work.

486	Observation	Prediction
487	0 1 7 4 0 7 0 2	0 1 7 4 0 7 0 2
488	1 2 4 1 2 7 5 9	1 2 4 1 2 7 5 9
489	3 1 6 2 8 4 2 4	3 1 6 2 8 4 2 4
490	4 0 6 1 4 8 7 5	4 0 6 1 4 8 7 5
491	4 8 4 2 4 3 9 9	4 8 4 2 4 3 9 9
492	4 7 6 9 4 9 3 2	4 7 6 9 4 9 3 2
493	9 8 5 2 0 9 2 1	9 8 5 2 0 9 2 1
	3 1 9 3 5 6 8 6	3 1 9 3 5 6 8 6

(a) Reconstructions

494	Class 0	Class 1
495	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
496	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
497	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
498	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
499	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
500	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
501	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
502	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
503	0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1
504	Class 2	Class 3
505	2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2
506	2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2
507	2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2
508	2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2
509	2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2
510	2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2
511	Class 4	Class 5
512	4 4 4 4 4 4 4 4	5 5 5 5 5 5 5 5
513	4 4 4 4 4 4 4 4	5 5 5 5 5 5 5 5
514	4 4 4 4 4 4 4 4	5 5 5 5 5 5 5 5
515	4 4 4 4 4 4 4 4	5 5 5 5 5 5 5 5
516	4 4 4 4 4 4 4 4	5 5 5 5 5 5 5 5
517	4 4 4 4 4 4 4 4	5 5 5 5 5 5 5 5
518	4 4 4 4 4 4 4 4	5 5 5 5 5 5 5 5
519	Class 6	Class 7
520	6 6 6 6 6 6 6 6	7 7 7 7 7 7 7 7
521	6 6 6 6 6 6 6 6	7 7 7 7 7 7 7 7
522	6 6 6 6 6 6 6 6	7 7 7 7 7 7 7 7
523	6 6 6 6 6 6 6 6	7 7 7 7 7 7 7 7
524	6 6 6 6 6 6 6 6	7 7 7 7 7 7 7 7
525	6 6 6 6 6 6 6 6	7 7 7 7 7 7 7 7
526	6 6 6 6 6 6 6 6	7 7 7 7 7 7 7 7
527	Class 8	Class 9
528	8 8 8 8 8 8 8 8	9 9 9 9 9 9 9 9
529	8 8 8 8 8 8 8 8	9 9 9 9 9 9 9 9
530	8 8 8 8 8 8 8 8	9 9 9 9 9 9 9 9
531	8 8 8 8 8 8 8 8	9 9 9 9 9 9 9 9
532	8 8 8 8 8 8 8 8	9 9 9 9 9 9 9 9
533	8 8 8 8 8 8 8 8	9 9 9 9 9 9 9 9

(b) Generative replay

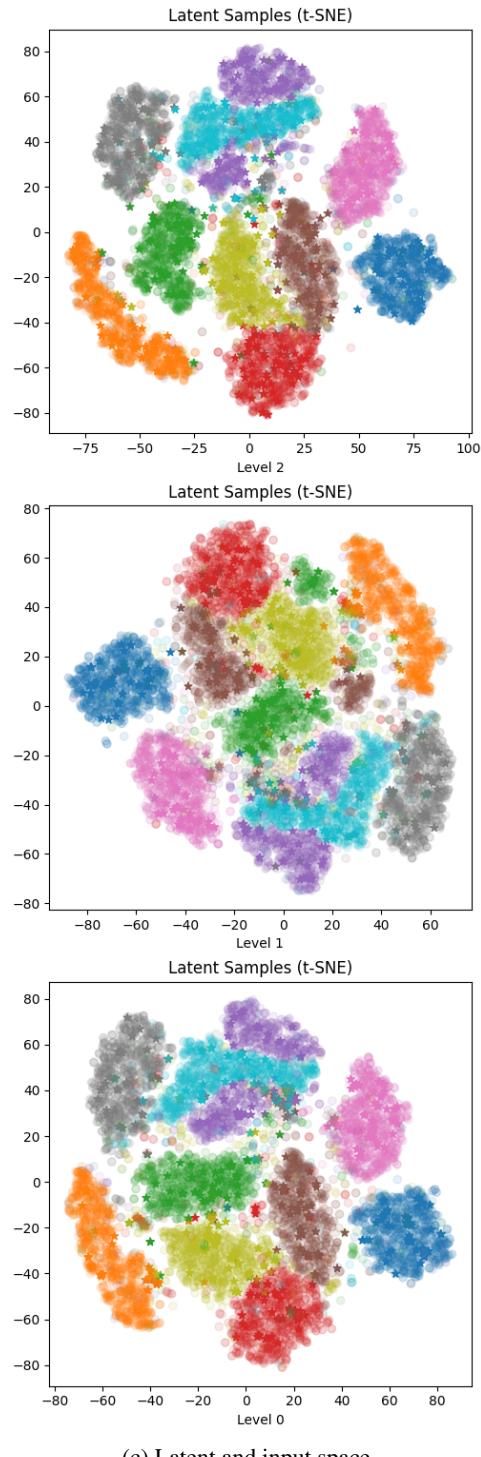


Figure 6: Visualizations for a model with $n_2 = 30$ units in the top level. (a) Images reconstructed (right) by a PCN with $n_2 = 30$ units in the top level for a random mini-batch of images from the validation set (left). (b) Images generated by replay for each class. (c) Hierarchical representations of the images generated by generative replay (star-shaped markers) and of the images of the validation set (transparent circle markers), visualized in 2D using t-SNE. Each image is represented as one data point in each of the three subplots, colored according to the class.

540 REFERENCES
541

542 Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature
543 of deep generative models, 2021. URL <https://arxiv.org/abs/1710.11379>.

544 Helen C Barron, Ryszard Auksztulewicz, and Karl Friston. Prediction and memory: A predictive
545 coding account. *Progress in neurobiology*, 192:101821, 2020.

546

547 Matthias Brucklacher, Sander M. Bohté, Jorge F. Mejias, and Cyriel M. A. Pen-
548 nartz. Local minimization of prediction errors drives learning of invariant object
549 representations in a generative network model of visual perception. *Frontiers in*
550 *Computational Neuroscience*, Volume 17 - 2023, 2023. ISSN 1662-5188. doi:
551 [10.3389/fncom.2023.1207361](https://doi.org/10.3389/fncom.2023.1207361). URL <https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2023.1207361>.

552

553 György Buzsáki. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and
554 planning. *Hippocampus*, 25(10):1073–1188, 2015. doi: <https://doi.org/10.1002/hipo.22488>. URL
555 <https://onlinelibrary.wiley.com/doi/abs/10.1002/hipo.22488>.

556

557 N. Alex Cayco-Gajic and R. Angus Silver. Re-evaluating circuit mechanisms underlying pattern
558 separation. *Neuron*, 101(4):584–602, 2019. ISSN 0896-6273. doi: <https://doi.org/10.1016/j.neuron.2019.01.044>. URL <https://www.sciencedirect.com/science/article/pii/S0896627319300716>.

559

560

561 Andy Clark. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University
562 Press, 01 2016. ISBN 9780190217013. doi: 10.1093/acprof:oso/9780190217013.001.0001. URL
563 <https://doi.org/10.1093/acprof:oso/9780190217013.001.0001>.

564

565 Harriet Feldman and Karl Friston. Attention, uncertainty, and free-energy. *Frontiers in*
566 *Human Neuroscience*, 4, 2010. ISSN 1662-5161. doi: 10.3389/fnhum.2010.00215.
567 URL <https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2010.00215>.

568

569 Lucie Fontaine and Frédéric Alexandre. Semantic and episodic memories in a predictive coding
570 model of the neocortex. In *2025 International Joint Conference on Neural Networks*, Rome,
571 Italy, June 2025a. URL <https://hal.science/hal-05202850>.

572

573 Lucie Fontaine and Frédéric Alexandre. Semantic and episodic memories in a predictive coding
574 model of the neocortex. In *Proceedings of the IEEE Int. Joint Conf. on Neural Networks 2025*,
575 Rome, Italy, June 2025b. URL <https://hal.science/hal-05202850>.

576

577 Karl Friston. Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352, 2003.

578

579 Jeff Hawkins, Marcus Lewis, Mirko Klukas, Scott Purdy, and Subutai Ahmad. A frame-
580 work for intelligence and cortical function based on grid cells in the neocortex. *Frontiers*
581 *in Neural Circuits*, Volume 12 - 2018, 2019. ISSN 1662-5110. doi: 10.3389/fncir.2018.
582 00121. URL <https://www.frontiersin.org/journals/neural-circuits/articles/10.3389/fncir.2018.00121>.

583

584 Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances*
585 *in neural information processing systems*, 29, 2016.

586

587 Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent
588 agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20
(7):512–534, 2016.

589

590 T. Ed Li. *Predictive Processing: Commitments, Unifying Mental Functions, and Perceptual Illu-
591 sions*. PhD thesis, University of Oxford, 2023.

592

593 T. Ed Li, Mufeng Tang, and Rafal Bogacz. Predictive coding model detects novelty on different
594 levels of representation hierarchy. *Neural Computation*, 37(8):1373–1408, 07 2025. ISSN 0899-
7667. doi: 10.1162/neco_a_01769. URL https://doi.org/10.1162/neco_a_01769.

594 Beren Millidge. Implementing predictive processing and active inference: Preliminary steps and
 595 results. March 2019. doi: 10.31234/osf.io/4hb58. URL <http://dx.doi.org/10.31234/osf.io/4hb58>.

597 Gaspard Oliviers, Rafal Bogacz, and Alexander Meulemans. Learning probability distributions
 598 of sensory inputs with monte carlo predictive coding. *PLOS Computational Biology*, 20(10):
 599 1–34, 10 2024. doi: 10.1371/journal.pcbi.1012532. URL <https://doi.org/10.1371/journal.pcbi.1012532>.

600 Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by
 601 learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

602 Alexander Ororbia and Daniel Kifer. The neural coding framework for learning generative models.
 603 *Nature Communications*, 13:2064, 04 2022. doi: 10.1038/s41467-022-29632-7.

604 Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpreta-
 605 tion of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

606 Tommaso Salvatori, Yuhang Song, Yujian Hong, Lei Sha, Simon Frieder, Zhenghua Xu, Rafal Bo-
 607 gacz, and Thomas Lukasiewicz. Associative memories via predictive coding. *Advances in Neural*
 608 *Information Processing Systems*, 34:3874–3886, 2021.

609 Eleanor Spens and Neil Burgess. A generative model of memory construction and consolidation.
 610 *Nature Human Behaviour*, pp. 1–18, 2024.

611 Ivilin Stoianov, Domenico Maisto, and Giovanni Pezzulo. The hippocampal formation as a hi-
 612 erarchical generative model supporting generative replay and continual learning. *Progress in*
 613 *Neurobiology*, 217:102329, 2022. ISSN 0301-0082. doi: <https://doi.org/10.1016/j.pneurobio.2022.102329>. URL <https://www.sciencedirect.com/science/article/pii/S0301008222001150>.

614 Mufeng Tang, Tommaso Salvatori, Beren Millidge, Yuhang Song, Thomas Lukasiewicz, and Rafal
 615 Bogacz. Recurrent predictive coding models for associative memory employing covariance learn-
 616 ing. *PLOS Computational Biology*, 19(4):1–27, 04 2023. doi: 10.1371/journal.pcbi.1010719.
 617 URL <https://doi.org/10.1371/journal.pcbi.1010719>.

618 Gido M. van de Ven, Hava T. Siegelmann, and Andreas Savas Tolias. Brain-inspired replay for
 619 continual learning with artificial neural networks. *Nature Communications*, 11, 2020. URL
 620 <https://api.semanticscholar.org/CorpusID:22111120>.

621 Renhao Wang, Kevin Frans, Pieter Abbeel, Sergey Levine, and Alexei A. Efros. Prioritized genera-
 622 tive replay, 2025. URL <https://arxiv.org/abs/2410.18082>.

623

632 A APPENDIX

633 A.1 ALGORITHM

634 The algorithm described in section 4.2 is summarized in algorithm 1 where α, β are the inference
 635 and learning rates. The constant $\frac{N}{M}$ from equation 7 is factorized in the learning rate β . To prevent
 636 instabilities which occurred systematically during training, we propose an exponential learning rate
 637 scheduler

$$\beta_{\text{epoch}} = \gamma \times \beta_{\text{epoch}-1}$$

638 which decays the learning rate β by a multiplicative factor γ at each epoch.

639 Initialization parameters include the standard deviations σ_W and σ_ϕ and the number of dimensions
 640 of latent state ϕ_0 .

641 A.2 EXPERIMENTS

648
649
650
651

Algorithm 1 Model training with minibatches

```

653  $W_1, W_2 \leftarrow$  Sample from  $\mathcal{N}(0, \sigma_W)$ 
654  $b_1 \leftarrow \mathcal{U}(-\frac{1}{n_0}, \frac{1}{n_0})$ 
655 repeat
656   for  $k = 1$  to  $n_{\text{batches}}$  do
657      $X^M \leftarrow$  Random minibatch of  $M$  datapoints drawn from  $X$ 
658     for  $i = 1$  to  $M$  do
659        $\phi_2^{(i)} \leftarrow x^{(i)}$ 
660        $\phi_0^{(i)}, \phi_1^{(i)} \leftarrow$  Sample from  $\mathcal{N}(0, \sigma_\phi)$ 
661        $\xi_0^{(i)}, \xi_1^{(i)}, \xi_2^{(i)} \leftarrow$  Calculate the corresponding errors (equation 4)
662       for  $t = 1$  to  $T$  do
663          $\phi_1^{(i)} \leftarrow \phi_1^{(i)} + \alpha(W_2^\top \xi_2^{(i)} - \xi_1^{(i)})$ 
664          $\phi_0^{(i)} \leftarrow \phi_0^{(i)} + \alpha(W_1^\top \text{diag} [f'(W_1 \phi_0^{(i)} + b_1)] \xi_1^{(i)} - \xi_0^{(i)})$ 
665          $\xi_0^{(i)}, \xi_1^{(i)}, \xi_2^{(i)} \leftarrow$  Calculate the corresponding errors (equation 4)
666       end for
667     end for
668      $W_2 \leftarrow W_2 + \beta \sum_{i=1}^M \xi_2^{(i)} \phi_1^{(i)\top}$ 
669      $W_1 \leftarrow W_1 + \beta \sum_{i=1}^M [\xi_1^{(i)} \odot f'(W_1 \phi_0^{(i)} + b_1)] \phi_0^{(i)\top}$ 
670      $b_1 \leftarrow b_1 + \beta \sum_{i=1}^M \xi_1^{(i)} \odot f'(W_1 \phi_0^{(i)} + b_1)$ 
671   end for
672 until variational free energy  $\mathcal{L}(\theta; X)$  is minimized
673

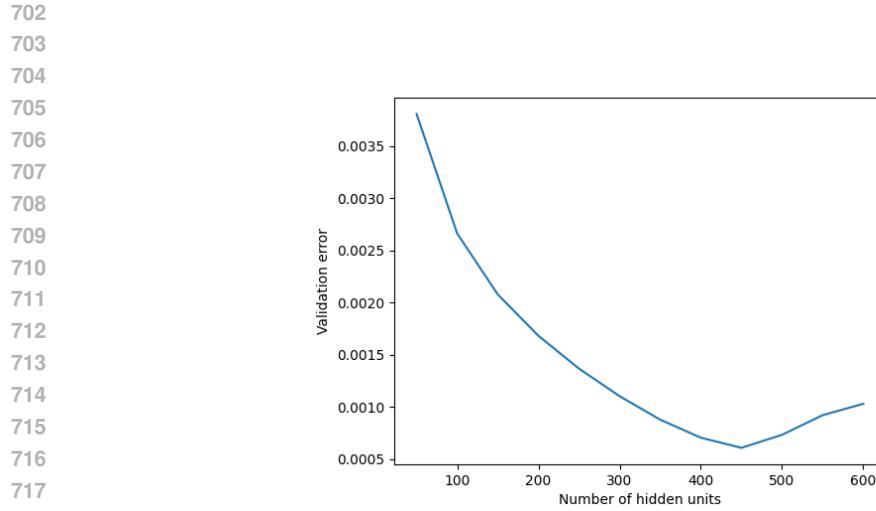
```

675
676
677
678
679
680

Table 1: Hyperparameter values

684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701

Parameter	Value
Activation function f	tanh
Batch size	64
Standard deviation σ_W	0.01
Standard deviation σ_ϕ	0.05
Number of iterations T_{train}	50
Number of iterations T_{valid}	200
Maximum number of iterations T_{max}	20000
Convergence threshold ϵ	2×10^{-4}
Inference rate α	0.01
Inference optimizer	SGD
Initial learning rate β_0	10^{-5}
Learning rate decay factor γ	0.99
Learning optimizer	Adam



719 Figure 7: RMSE between original and reconstructed images averaged over the validation set for a
720 PCN with $L = 1$.

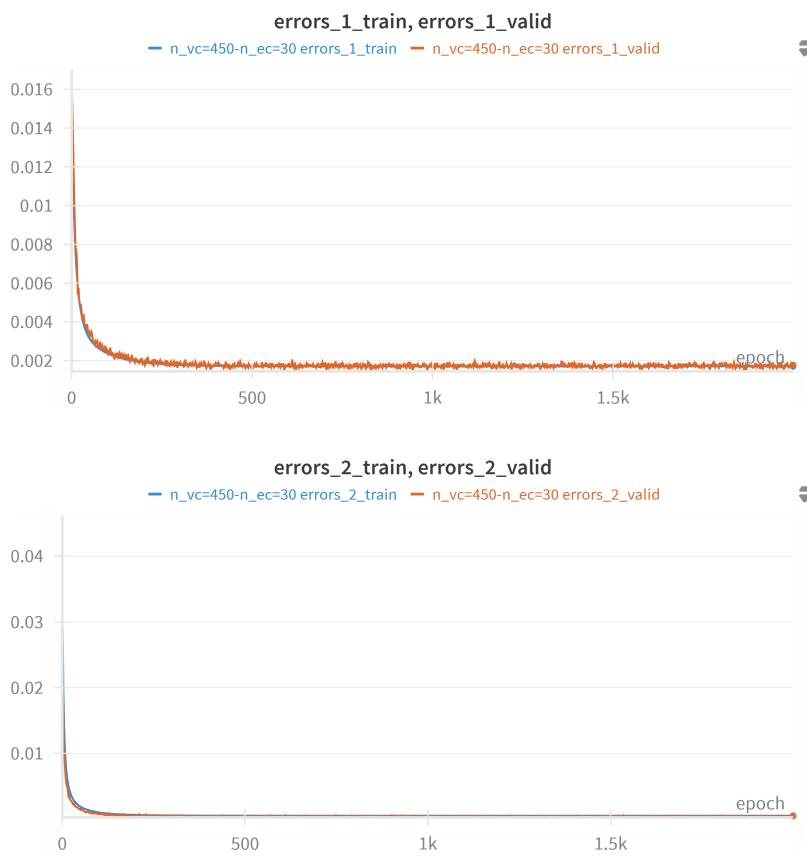


Figure 8: Mean prediction errors in level 1 and 2 of a model with 30 units in level 0, averaged over the whole training set (blue line) and a random minibatch from the validation set (orange line) at each epoch. The model converges after 2000 epochs thanks to the learning rate scheduler.